LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

LLNL-CONF-795318

Selecting Parameters for Image Processing Algorithms: A Case Study Using Retinal Image Segmentation

C. Kamath

October 23, 2019

Asilomar Conference on Signals, Systems, and Computers
Pacific Grove, CA, United States
November 3, 2019 through November 6, 2019

Disclaimer

This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

Selecting Parameters for Image Processing Algorithms: A Case Study Using Retinal Image Segmentation

Chandrika Kamath
Center for Applied Scientific Computing, Lawrence Livermore National Laboratory
Livermore, CA
kamath2@llnl.gov

***Abstract*—A challenge in applying image processing algorithms to scientific data is the variation within and across images in a data set. This makes it difficult to select parameters for the algorithms, especially when the data set is large and the variation is unknown. Using the task of segmentation of retinal images, we discuss the challenges that arise in selecting parameters for even a relatively simple algorithm. We show that the availability of ground truth results could make the identification of parameters subjective and propose a simple idea that could benefit the selection of a single parameter.**

***Index Terms*—image segmentation, parameter selection**

I. Introduction

Selecting parameters for image processing algorithms is often done by trial and error. As algorithms become more complex, and the number of parameters increases, finding an optimal set of parameters is a challenge. This is exacerbated when there is a large variation both within and across images in a data set, and the size of the data set precludes a visual inspection of the results. Processing a small subset of images to generate ground-truth results for comparison can be quite helpful, but generating such a set is often tedious and the results are subjective. We are therefore interested in alternate ideas for correctly processing large-scale image data sets.

We conduct our study using the public-domain, STARE data set [1], which is a collection of 400 retinal images that were used to develop automated methods to locate and outline blood vessels in the images. Ground truth results are available for twenty images, generated by two researchers, "AH" and "VK".

Figure 1 shows two images — one normal and one diseased. There is a large variation in intensity within each image, with several bright and dark regions. The intensity along a blood vessel is not always uniform and a vessel may have both bright and faint parts. Blood vessels can be quite thin and faint, like a filament, or wide and bright. Other nearby blood vessels may also make it difficult to segment the image correctly.

This data set has been analyzed extensively, using methods ranging from traditional image processing to classification algorithms and even deep learning to directly map the image to the associated disease. It is not our intent to generate yet another algorithm for segmenting these images. However, the variation in the images, the availability of the ground truth, and the relatively large size of the data set makes this an ideal test bed to understand how we might select the parameters for typical image processing algorithms.

This work is part of the IDEALS project, funded by ASCR (program managers, Dr. Nowell and Dr. Biven), U.S. Department of Energy.

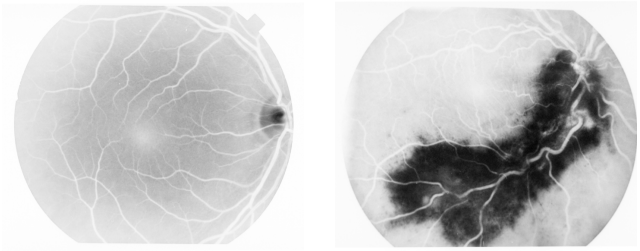

Fig. 1. The green band of two images, one normal (left, im0082) and one with retinitis (right, im0044), from the STARE data set. The green channel has the highest contrast between blood vessels and other retinal structures. We show the negative of the images for clarity.

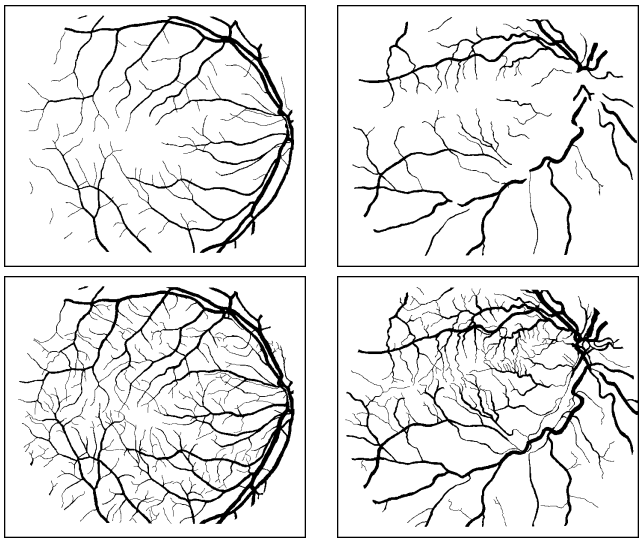

Fig. 2. The ground truth images corresponding to the two images in Figure 1. The top row was generated by researcher "AH" and the bottom row by researcher "VK". A black border has been added to each image for clarity.

Figure 2 shows the two ground truth images associated with each image in Figure 1. The two sets of images are quite different, as one identifies several thin blood vessels that are

not considered in the other. We also observe that humans make assumptions in generating the ground truth, such as identifying pixels in a faint region as vessel pixels as they are the continuation of pixels in a vessel in a brighter region nearby. However, such human knowledge is difficult to encode into an algorithm, which may limit the accuracy obtained by the algorithm. Therefore, we need to be careful in using ground truth results to evaluate algorithms or to set parameters.

II. Proposed method

In our study, we use a simple gradient-based algorithm to detect the blood vessels in the images. We start by smoothing the image using a Gaussian filter of width 7 that is applied twice to the image. This reduces the noise in the image and makes it easier to select parameters for subsequent steps. Next, using the definition of gradient from the work of Horev et al. [2], we calculate the vertical gradient at a location (x, y) in the image I as follows:

$$\begin{aligned} g_y(x,y) =& I(x,y+3) + I(x,y+2) + I(x,y+1) - \\ & I(x,y) - I(x,y-1) - I(x,y-2), \end{aligned}$$

and the horizontal gradient as

$$\begin{aligned} g_x(x,y) =& I(x+3,y) + I(x+2,y) + I(x+1,y) - \\ & I(x,y) - I(x-1,y) - I(x-2,y) \end{aligned}$$

Other authors have proposed calculating the gradient at several angles [3], but we found that using just two gradients was not only adequate, but also reduces the compute time and memory requirements. We then observed that in the region of a blood vessel, the vertical gradient has a low, negative value on the top part of the vessel, and a high, positive value below, while the horizontal gradient has a high, positive value on the left and a low, negative value on the right. So we can identify the location of a vessel by searching for peaks and valleys that are adjacent, with the peak greater than a threshold, thr_g, and the valley less than a threshold, $-thr_g$. Typically, these peaks and valleys were separated by a small number of pixels, which we set to 1; this finds the skeleton of the vessel. We then use region growing and morphological operations to identify pixels that belong to the wider blood vessels and to smooth their boundaries. Our approach is a much simplified version of the one by Mendonca et al. [3] that combines detection of centerlines with morphological region growing.

The segmentation of an image results in assigning a label to each pixel to indicate if it is part of a blood vessel or the background. We can then use the ground truth images to create a confusion matrix (Table I) that counts the number of pixels that are correctly or incorrectly identified. For example, TP (true positive) is the number of blood vessel pixels that are correctly identified, while FN (false negative) is the number of blood vessel pixels identified falsely as a background pixel.

Next, using the values in the confusion matrix, we can generate various metrics [4] that indicate the quality of segmentation as defined in Table II. For example, accuracy reflects the fraction of pixels correctly identified as either blood vessel pixels or background pixels, while sensitivity or recall is the fraction of the blood vessel pixels that were correctly identified as such, and precision is the fraction of true blood vessel pixels among those labeled as such in the segmentation.

TABLE I
Confusion matrix used to generate metrics for segmentation quality

	Predicted positive	Predicted negative
Actual positive	True positive (TP)	False negative (FN)
Actual negative	False positive (TP)	True negative (tN)

TABLE II
Segmentation quality metrics

Metric	Formula
Accuracy	(TP+TN)/(TP+TN+FN+FP)
Error	(FP+FN)/(TP+TN+FN+FP)
Sensitivity (= Recall REC)	TP/(TP+FN)
Specificity	TN/(TN+FP)
False positive rate (FPR)	FP/(TN+FP)
False discovery rate (FDR)	FP/(TP+FP)
Precision (PREC)	TP/(TP+FP)
Matthews Corr. Coeff.	(TP*TN-FP*FN)/ ((TP+FP)(TP+FN)(TN+FP)(TN+FN))$^{0.5}$
$F_{0.5}$	1.5 * PREC * REC /(0.25 * PREC + REC)
F_1	2.0 * PREC * REC /(PREC + REC)
F_2	5.0 * PREC * REC /(4.0 * PREC + REC)

Note that a metric reflects a specific aspect of the quality of segmentation and we may need more than one metric to evaluate the quality of segmentation for a given value of thr_g. For example, we want to correctly identify the blood vessel pixels (=TP), but also want to reduce the number of pixels incorrectly identified as blood-vessel pixels (=FP). Further, it is well known that for unbalanced data sets, such as ours, with far more background pixels than blood-vessel pixels, some error metrics, such as accuracy can give misleading results.

In our problem, we need to determine one parameter, the gradient threshold, thr_g, for segmenting the 20 test images. Each test image has two associated ground truth images. We want to address three questions:

- Which metric, derived from the ground truth, should we use to determine thr_g?
- Does the subjectivity in the two ground truth images have an effect on the choice of thr_g?
- Is there any other metric we can use to determine thr_g?

To address the last question, we observed that the gradient threshold, thr_g, determines which pixels in the image are considered to be blood vessel pixels as they have a high contrast or gradient with the background. When thr_g is very small, many objects, that is, connected pixels that meet the threshold criterion, are created. Some of these objects reflect the noise in the images. As thr_g is increased, the number of objects reduces, until just the brightest blood vessels, with the highest contrast, are detected. Since we also want to detect the fainter blood vessels, a moderate threshold seemed

appropriate. This prompted us to consider the number of objects detected in the image as another metric for selecting thr_g. Note that this metric does not depend on the availability of the ground truth images.

III. Experimental Data and results

To address the three questions posed in Section II, we first segmented the 20 test images using a set of thresholds in the range [0.0001,0.009]. Figure 3 shows the segmentation corresponding to the test images in Figure 1 for three different thresholds — 0.0001 (the smallest threshold), the optimal threshold for each image (identified later in this section), and 0.009 (the largest threshold). We observe that the very small threshold identifies several spurious blood vessels outside the retina, while the very large threshold misses several obvious blood vessels. This confirms that we have selected a reasonably large range of threshold values for our experimentation.

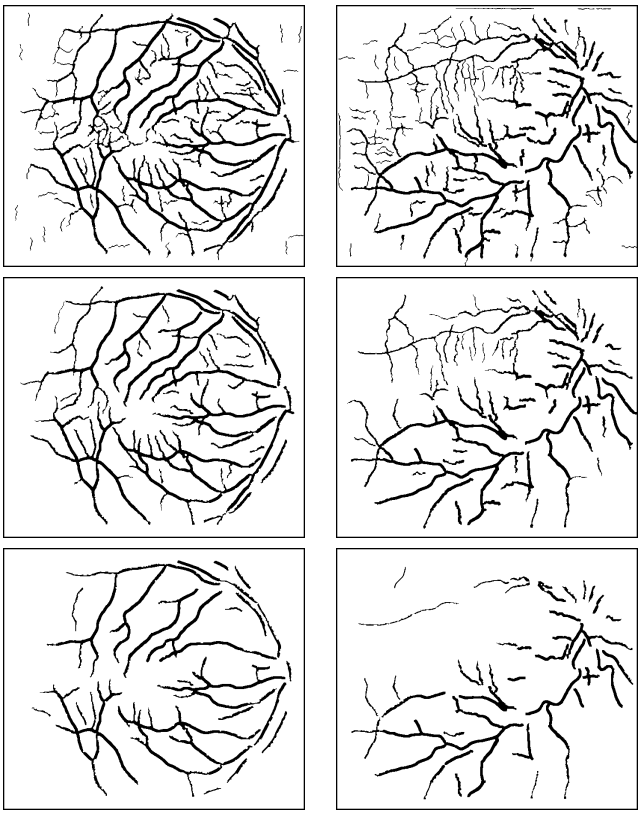

Fig. 3. The blood vessels identified in the two images in Figure 1 — im0082 (left) and im0044 (right) — as a function of three different gradient thresholds. Top: 0.0001; middle: 0.0025 (left) and 0.002 (right); bottom: 0.009.

We then compared the segmented images with the two ground truth images ("AH" and "VK") and generated the metrics in Table II, which are presented in Figure 4 for the first ten test images. For comparison, the range of values of sensitivity, specificity, and accuracy from previous studies is [0.67,0.75], [0.96,0.98] and [0.90,0.96], respectively, across all 20 images in the data set [5]. We make two observations:

- First, the results with the two ground truth images tend to be similar, in terms of the shapes of the curves, though the values of the metrics obtained for each are different. One difference we note is with the accuracy plots (panels (e)) where the black dotted curve for image 139 has a different shape between the two ground truth images. This suggests that even though the two sets of ground truth images are clearly different, they are not different enough for many of the images to change the shape of the curves. That is, the fine structures identified in the "VK" images appear not to influence the choice of threshold.
- Second, it is unclear from these plots how we should select the optimal threshold to use for an image. For example, the accuracy appears to increase with increasing thr_g, when only the very bright blood-vessel pixels are detected. In fact, it is the resulting increase in the large number of background pixels that improves the accuracy rather than the correct detection of the blood vessels. Further, for the three F-scores, the values are sometimes nearly constant over a range of thresholds, making it unclear which threshold value is optimal.

We next considered two commonly used curves to determine the optimal threshold - the precision-recall curve and the ROC curve that relates sensitivity to the FPR, which is (1.0-specificity). Though ROC curves are frequently used to determine thresholds, they are considered to be less accurate for unbalanced data sets, such as our test images [4]. Figure 5 shows these curves for all twenty test images, in two sets of ten each, for the two ground truth images. Typically, the threshold is selected as the knee of these curves as we want both a high precision and a high recall, if using the precision-recall curve, or a high true positive rate with a low false positive rate, if using the ROC curve.

Figure 6 shows our proposed metric, the number of objects detected in the images, as a function of the gradient threshold. The values of the threshold at the knee of the curves in Figures 5 and 6 are summarized in Table III. This table indicates that there is little difference between the thr_g values derived from the two ground truth images. One data point that stands out is im0005, where the ROC curve gives different optimal threshold values (0.004 and 0.006). The ground truth images for this test image (Figure 7) are quite different as "VK" identifies large blood vessels near the center of the image, while "AH" does not. Interestingly, the precision-recall curve does not seem affected by this difference, though the identical threshold value of 0.006 may be a chance coincidence.

More interestingly, when we examine the results for im0005 at the the three different optimal values of thr_g — 0.0025, 0.004, and 0.006 — identified in Table III, we find that the images are very similar. This suggests that there is a range of gradient threshold values over which the segmentation does not change by a large amount. We also observe that the threshold at the knee of the number of objects curve is nearly the same across the different images, though the images themselves are quite different. This suggests that our algorithm for segmenting the retinal images is robust to the choice of the gradient threshold, which can be chosen independent of the ground truth images and the variation in the images.

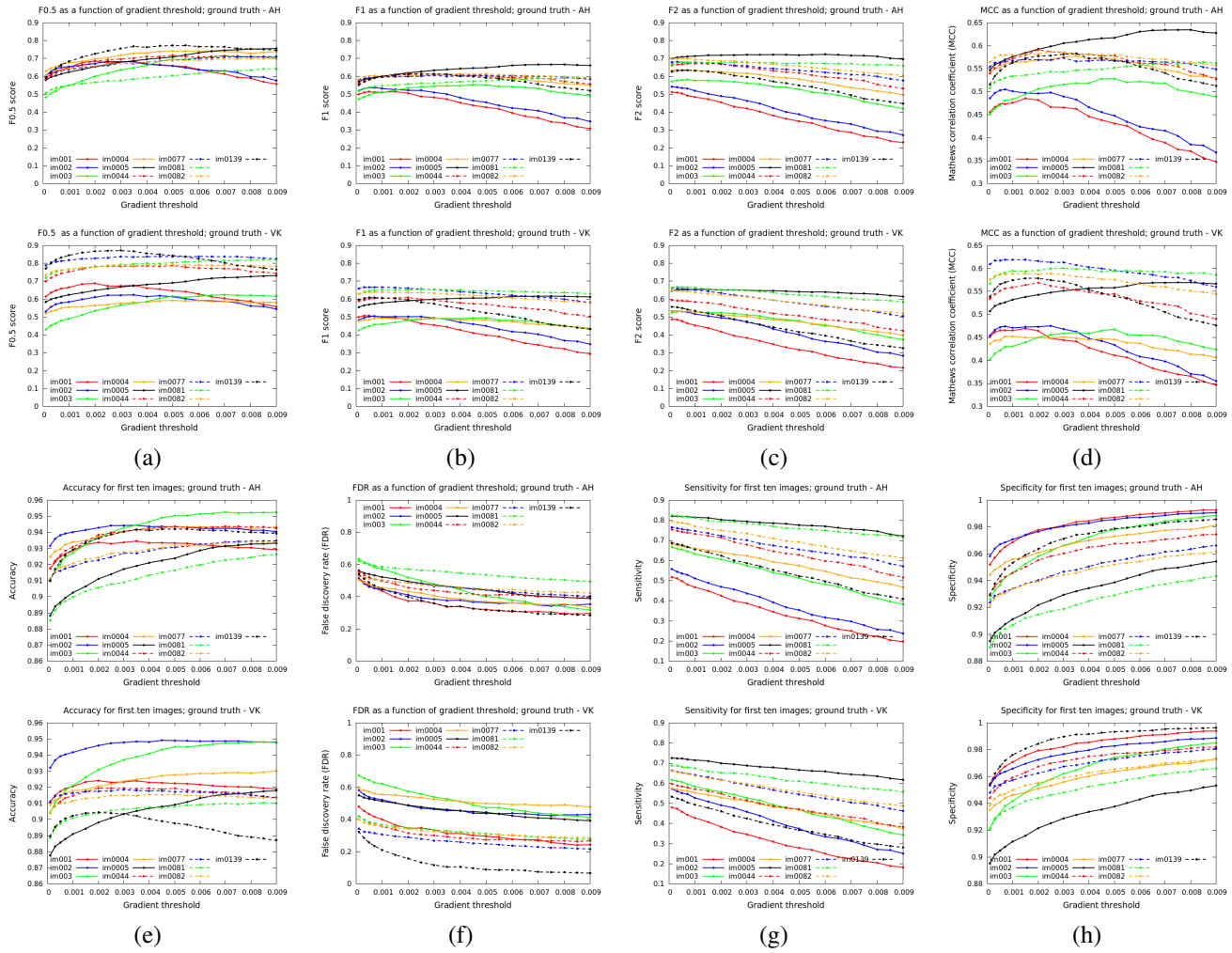

Fig. 4. Quality metrics for the first ten test images as the gradient threshold is varied. Rows 1 and 3 are with the "AH" image and rows 2 and 4 are with the "VK" image. Rows 1 and 2, from left to right: $F_{0.5}$, F_1, F_2, and MCC. Rows 3 and 4, from left to right, accuracy, FDR, sensitivity, and specificity.

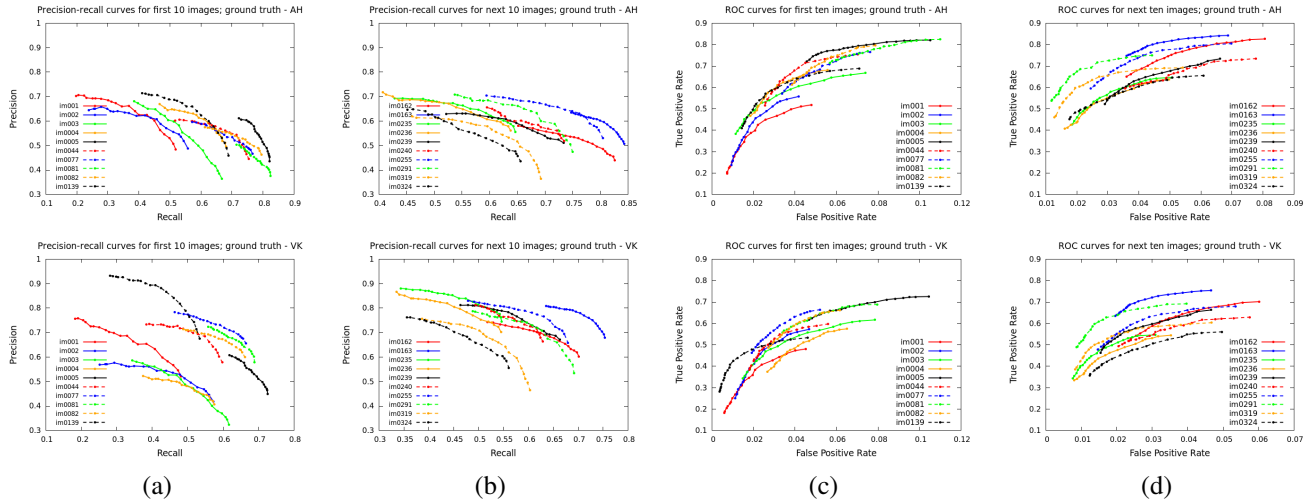

Fig. 5. The precision-recall curves ((a) and (b)) and the ROC curves ((c) and (d)) for all twenty test images, in two groups of ten each, using ground truth "AH" (top row) and "VK" (bottom row) images.

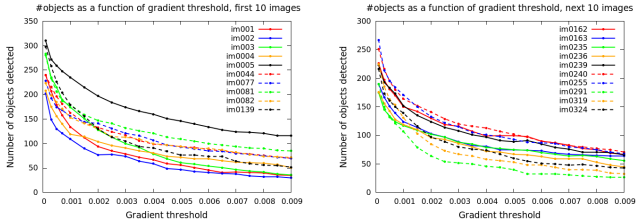

Fig. 6. The variation in number of objects detected in the vertical and horizontal gradient images for the first ten images (left) and next ten images (right).

TABLE III
OPTIMAL THRESHOLD IDENTIFIED AS THE KNEE OF VARIOUS CURVES.

Image	ROC-AH	ROC-VK	PR-AH	PR-VK	NumObjs
im0001	0.002	0.002	0.002	0.002	0.002
im0002	0.0025	0.0025	0.0025	0.0025	0.002
im0003	0.003	0.003	0.0045	0.005	0.0025
im0004	0.0035	0.003	0.0035	0.0045	0.001
im0005	0.006	0.004	0.006	0.006	0.0025
im0044	0.002	0.002	0.002	0.002	0.002
im0077	0.003	0.003	0.003	0.003	0.0015
im0081	0.0025	0.0025	0.001	0.0025	0.002
im0082	0.0025	0.0025	0.0025	0.0025	0.0025
im0139	0.003	0.002	0.0035	0.003	0.0025
im0162	0.004	0.0025	0.004	0.004	0.0025
im0163	0.0035	0.002	0.0035	0.0035	0.002
im0235	0.003	0.003	0.003	0.003	0.003
im0236	0.004	0.003	0.004	0.004	0.002
im0239	0.0045	0.0045	0.0045	0.0045	0.0025
im0240	0.001	0.001	0.001	0.001	0.0025
im0255	0.002	0.002	0.002	0.002	0.002
im0291	0.0025	0.0025	0.0025	0.0025	0.0025
im0319	0.0025	0.0025	0.005	0.0025	0.0025
im0324	0.002	0.0035	0.002	0.0035	0.002

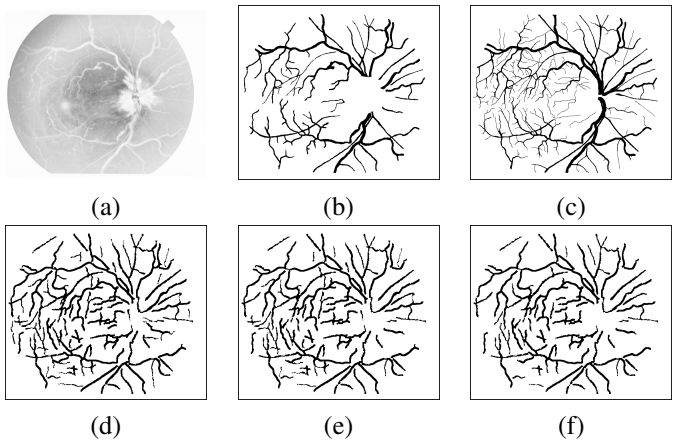

Fig. 7. The blood vessels identified in image im0005. (a) original image (b) ground truth "AH", (c) ground truth "VK" that includes large blood vessels near the center, (d) thr_g = 0.0025, (e) thr_g = 0.004, and (f) thr_g = 0.006. The gradient thresholds are the optimal values from Table III.

IV. CONCLUSION

In this paper, we considered the problem of selecting parameters for an image segmentation algorithm. Using a test bed of twenty retinal images and two ground truth images for each, we investigated different metrics we could use to select one parameter, a gradient threshold, used in our segmentation algorithm. We found that even though the different metrics and ground truth images gave different values for the optimal threshold, these values were close enough and in a region where the algorithm was relatively insensitive to the value of the threshold. For this specific test bed, it appears that the subjective differences in the ground truth images has little influence on the choice of the parameter we were varying. A simple metric that is independent of ground truth images also gave good results, though we expect that designing an algorithm robust to the choice of parameters is more important.

ACKNOWLEDGMENT

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. We thank the authors of [1] for making their data set available for use.

This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes. LLNL-CONF-795318.

REFERENCES

[1] A. D. Hoover, V. Kouznetsova, and M. Goldbaum, "Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response," *IEEE Transactions on Medical Imaging*, vol. 19, no. 3, pp. 203–210, March 2000.

[2] I. Horev, B. Nadler, E. Arias-Castro, M. Galun, and R. Basri, "Detection of long edges on a computational budget: A sublinear approach," *SIAM Journal on Imaging Sciences*, vol. 8, pp. 458–483, 2015.

[3] A. M. Mendonca and A. Campilho, "Segmentation of retinal blood vessels by combining the detection of centerlines and morphological reconstruction," *IEEE Transactions on Medical Imaging*, vol. 25, no. 9, pp. 1200–1213, Sep. 2006.

[4] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," *PLOS ONE*, vol. 10, no. 3, pp. 1–21, March 2015. [Online]. Available: https://doi.org/10.1371/journal.pone.0118432

[5] M. M. Fraz, P. Remagnino, A. Hoppe, B. Uyyanonvara, A. R. Rudnicka, C. G. Owen, and S. A. Barman, "An ensemble classification-based approach applied to retinal blood vessel segmentation," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 9, pp. 2538–2548, Sep. 2012.